THE 1996 BROADCAST NEWS SPEECH AND LANGUAGE-MODEL CORPUS

David Graff

Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA

ABSTRACT

The Linguistic Data Consortium handled the recording, digitization, and transcription of 130 hours of radio and television news broadcasts. Of this material, 50 hours' worth was designated and published as a baseline training set, and 20 hours were prepared and distributed by NIST for use as development and evaluation test data, for the 1996 ARPA CSR Benchmark Tests. The remaining 60 hours were held in reserve for use as additional training and evaluation test data in the 1997 benchmarks. The LDC also acquired, conditioned and published a five-year archive of commercially produced broadcast transcripts for use in constructing language models for the broadcast news domain. These tasks posed a broad range of novel challenges for the LDC, as well as for those at NIST and elsewhere who were involved in defining, clarifying and applying the standards and requirements for this corpus. This paper summarizes the content of the corpus, reviews the established specifications for the transcription format, and briefly describes the tools and methods used in the transcription effort.

1. INTRODUCTION

The November 1996 Benchmark Test for the ARPA Continuous Speech Recognition program (CSR), also referred to as the 1996 HUB-4 Evaluation, represents the first attempt in this program to focus entirely on "found speech" – that is, on speech that has been observed and captured in actual day-to-day usage – in contrast to the benchmark tests of previous years, which were based on speech elicited solely for purposes of speech recognition research.

In order to provide some continuity with previous years, in which the speech consisted of readings of journalistic text, the HUB-4 corpus focuses on news broadcasts over radio and television networks. In particular, recordings were made from broadcasts by the ABC, CNN and CSPAN television networks, as well as the NPR radio network.

The Linguistic Data Consortium was assigned the task to provide the recordings and transcriptions for use in acoustic training and tests, as well as a suitable body of related text data to support language model training. The goals for this task were as follows:

- Negotiate with networks to acquire permissions for recording and redistributing both audio and video portions of broadcasts, as well as text transcriptions.
- Record and distribute 100 hours of radio and television broadcasts for use as training data, plus an additional 30 hours of broadcasts for use as development and evaluation test material.
- Coordinate with HUB-4 participants to implement a suitable specification for transcriptions of audio data.
- Transcribe all audio recordings in accordance with the specification.
- Acquire a multi-year archive of commercially produced broadcast transcripts, condition the collection in accordance with established conventions for CSR language model development, and distribute the finished form to HUB-4 participants.

Each of these tasks involved some novel activities for the LDC, and a couple of them were found to be much more complex and ambitious than expected, to the extent that they proved to be unattainable within the allotted time. The full and final definition of the transcript specification, which involved considerable input from committees and working groups outside the LDC, was not completed until long after the transcription effort was underway, while the unforeseen difficulty of transcription (amplified by the problems in finalizing the specification after transcription had begun) resulted in our falling short by half in the delivery of usable acoustic training transcripts in time for the November evaluation – only 50 hours of training data were made available, and the distribution of this material was completed only one month before the evaluation was to begin.

In its final form, released by the LDC in February 1997, the training corpus contained a total of 104 hours of broadcast recordings, with transcripts. In addition, there are three sets of 10 hours each from time periods subsequent to that of the training corpus. The first of these was designated as the development test set for the 1996 benchmark, the second was the evaluation test set for 1996, and the third is for use in the 1997 evaluation.

2. DESCRIPTION OF THE DATA COLLECTION

2.1. Data Sources

The LDC established contracts with the ABC, CNN and CSPAN television networks, and with National Public Radio, allowing us to record and redistribute a variety of news programs for research purposes. In addition, we obtained permission from Primary Source Media (PSM), a commercial distributor of broadcast transcripts, to condition and redistribute a four-year archive of transcript texts for use in language model development, drawing from their CD-ROM publications.

Tables 1 and 2 show the programs involved in the acoustic training and test collections, and the amount of material recorded. The time periods sampled for training data are shown in the first table; the sampling periods for test data were set as follows: July 10 to 15 for the development test set, September 11 to 25 for the 1996 evaluation test pool, and October 14 to November 13 for the 1997 evaluation

test pool. Table 3 summarizes the contents of the language model text collection drawn from PSM materials.

Network/Program	Date Range of Sample	Episodes/ Hours Recorded
ABC/Nightline	5/21 - 6/26	23/11.5
ABC/World Nightly News	5/29 - 6/20	15/13.0
ABC/World News Tonight	5/16 - 6/10	12/6.0
CNN/Early Edition	5/15 - 6/04	4/5.0
CNN/Early Prime News	5/10 - 5/22	9/8.5
CNN/Headline News	5/31 - 7/03	13/8.5
CNN/Prime News	5/14 - 6/11	12/6.0
CNN/The World Today	5/14 - 7/02	7/7.0
CSPAN/Washington Journal	5/31 - 6/11	7/14.0
NPR/All Things Considered	5/10 - 6/21	13/24.5
NPR/Marketplace	5/23 - 6/14	15/7.5
Total	5/10 - 7/03	130/111.5

Table 1: Summary of broadcast sampling for acoustic training data.

Network/Program	Hours recorded
ABC/Prime Time News	1
CNN/Morning News	2
CNN/World View	1
CSPAN/Washington Journal	2
NPR/Morning Edition	2
NPR/Marketplace	1
NPR/The World	1

Table 2: Contents of each acoustic test pool.

Network	Programs Represented	Story units
ABC	14	3320
CNN	60-80	27197
NPR	6	2524
PBS	20	976

Table 3: Summary of contents for language model text collection.

2.2. Recording Protocols

Television broadcasts were received via the campus-wide cable TV network at the University of Pennsylvania, and recorded simultaneously to both Super-VHS video tape and digital audio tape (DAT). A cable signal splitter and multiple video/audio recorders were used to allow recording from different networks simultaneously when necessary. Radio broadcasts were received by means of a common high-fidelity stereo receiver with digital FM tuner. An amplifying FM antenna was attached to the receiver and positioned within the LDC offices to maximize received signal strength. Distance to the broadcasting antennas was approximately 10 miles, well within the broadcasting range of the local NPR affiliate station (which reaches a radius of at least 60 miles).

The DAT recordings were played through a Townshend DATLink digital audio converter for downsampling from the 32 KHz DAT sample rate to 16 KHz and storage of the left channel only to 16-bit PCM encoded waveform sample files. In most cases where a single broadcast episode lasted for an hour or more, the waveform data were split into consecutive 30-minute segments for transcription and distribution. The files were formatted with NIST SPHERE headers, and arranged on CD-ROMS for publication.

Network/Program	Released in Oct. 96	Added in Feb. 97
ABC/Nightline	4.5 (3.01)	7.0 (4.74)
ABC/World Nightly News	4.5 (2.14)	8.0 (3.90)
ABC/World News Tonight	4.5 (3.02)	1.5 (1.07)
CNN/Early Edition	4.5 (2.78)	0.5 (0.21)
CNN/Early Prime News	3.5 (2.48)	5.0 (3.53)
CNN/Headline News	4.5 (2.84)	4.0 (2.58)
CNN/Prime News	4.5 (3.23)	1.0 (0.74)
CNN/The World Today	4.0 (2.63)	3.0 (2.00)
CSPAN/Washington Journal	4.0 (4.00)	8.0 (7.98)
NPR/All Things Considered	7.0 (5.74)	13.0 (9.24)
NPR/Marketplace	4.5 (3.63)	3.0 (2.23)
Total	50.0 (35.50)	54.0 (38.22)

Table 4: Total hours of recordings (and transcribed speech) in the training set

There were some irregularities in the recording schedule, involving unannounced delays by networks in their presentation of some programs. This resulted in some episodes starting later than expected in the recordings, and being cut short at the end of the recording period. (A number of ABC Nightline broadcasts were affected in this way, to the extent that five to ten minutes at the end of the broadcasts were missing from the tapes.)

2.3. Language Model Text Conditioning

The conditioning of the LM text collection was simplified to a large extent by the fact that most of the software needed for this task was already available. BBN provided a program they had developed to extract text data from the PSM CD-ROM publications, whose format was peculiar to a commercial text search engine provided on the PSM discs. Following this process, it was possible to apply the same conditioning that was used for previous CSR LM text collections – tagging of sentence boundaries and punctuation, conversion of abbreviations and digit strings to appropriate lexical tokens – with fairly minor adjustments to existing code and some additional preconditioning of the texts. (The method for sentence boundary detection was developed at the LDC, and the other conditioning steps were carried out using methods originally developed by Doug Paul for Wall Street Journal texts.)

2.4. Content and Release of Transcripts

While it was intended that transcriptions should be supplied to cover the full extent of each recorded broadcast, it was also agreed that some portions of broadcasts were not subject to transcription, either because they were judged as unsuitable to the research task (such as traffic reports or summaries of sports scores), or because the LDC had not obtained permissions from copyright holders to use the materials for research (such as commercials, or portions of other programs that immediately preceeded or followed the target broadcast). It

was found that different programs had different amounts of usable material. Table 4 indicates, for each program in the training corpus, the amounts of total recorded time and and total amount of transcribed speech (in hours); the table also indicates how the training corpus was partitioned to provide a balance of sources in the initial release of 50 hours, sent to researchers in October 1996.

3. TRANSCRIPTION FORMAT AND METHODS

3.1. Functional Requirements for Broadcast Transcriptions

Given the variety of speech and signal conditions observed in the broadcast recordings, together with the topical organization of broadcasts into independent stories, discussions and transitions, it was decided that the transcripts should contain a considerable amount of information that was qualitative, categorial, and structural in nature, to supplement the text of the recorded utterances. So, in addition to transcribing the speech, transcribers were also assigned the following tasks:

- Mark the beginning and ending of each topical unit and identify its type (report, transition, commercial, etc); some of these units (commercials, sports summaries, traffic reports) were to be left untranscribed, but their boundaries and types must still be marked.
- Mark the beginning and ending of each speaking turn (i.e. change of speaker) within a topical unit, identifying each speaker uniquely (by name if possible), and indicating gender and whether s/he is a native speaker of American English.
- For each speaking turn, indicate the channel quality (judged subjectively as high, medium or low fidelity), and speaking mode (spontaneous or planned).
- Mark the beginning and ending points of three types of background sound conditions (music, voices, varied noise), indicating the relative prominence of the condition (judged subjectively as low or high); when these conditions change in the middle of a speaking turn, mark the point of change at the nearest word boundary.
- When two speaking turns overlap in time (i.e. two speakers are talking simultaneously), mark the extent of text in each turn (at the nearest word boundary) that is affected by the overlap.

The judgements of channel quality were intended to reflect, roughly, recordings in a studio environment (high fidelity), in various field settings (medium), and in various band-limited conditions such as telephone interviews (low); in making these judgements, transcribers did not have access to any specialized signal processing tools. With regard to speaking mode, speech that was judged by the transcriber to have been read, prepared or formulaic in nature was identified as "planned"; speech that was evidently unscripted (and that typically contained disfluencies and/or hesitations) was marked as "spontaneous."

3.2. SGML Structure of Transcriptions

Owing to the complexity and hierarchical nature of the additional information needed in the transcripts, SGML was chosen as the most suitable framework to use in formatting the text. The document structure used for all transcripts is as follows:

- For each waveform file (whether a full program or a 30 minute portion of a longer program), there is one document (transcript) file, containing a single "Episode" element; the Episode has attributes to identify the file name, the transcriber, and the release version.
- Each Episode contains a series of "Section" elements, which equate to the topical units (stories, etc) in the Episode; the Section attributes identify the type of unit, and the points in time at which the Section begins and ends in the corresponding waveform file.
- Within each Section containing material to be transcribed, there are one or more "Segment" elements, corresponding to speaker turns within the Section; the Segment attributes identify the speaker, the speaking mode, the channel fidelity, and the points in time at which the speaking turn begin and end.
- At any point within an Episode, Section or Segment where there is a change in the presence or prominence of music, background voices or other noise, a "Background" element is inserted to mark the change; the Background attributes identify the type of background condition (music, voice, noise), the relative prominence of the condition following the change (high, low, off), and the point in time at which the change occurs.

In this design, a hierarchical structure exists among the Episode, Section and Segment elements, and this is reflected in the markup by requiring that these elements always have explicit end-tags, in addition to the latter two having start-time and end-time attributes, to define their extent in the transcript documents. The Background element is non-hierarchical, and is defined in the SGML Document Type Definition (DTD) as an empty element (having no embedded content); the sole purpose of this element in the document is to mark points in time at which background conditions change. An additional empty element, called "Sync", was defined as a convenience to transcribers; its sole attribute is a time value, and its purpose is to provide "break points" for auditing and display of waveform data – this was useful for breaking up long Segments into manageable pieces for transcription and checking.

Except for the Episode tags, which bound the entire transcript document, all SGML tags contain either a single time value (Background and Sync) or two end-point time values (Section and Segment). The placement of transcription text relative to the time-marked SGML tags was strictly constrained to coincide with the time sequence indicated by the tags. All words contained between the beginning and end tags of a Segment unit are to be heard when playing the portion of waveform data between the beginning and ending times for that Segment; in addition, text is placed before (or after) a Background or Sync tag *only* if the corresponding words are heard before (or after) the point in time marked by that tag.

The only condition for which the transcripts would violate the constraint of strict temporal linearity is the occurrence of overlapping speech. When consecutive Segments produced by two speakers overlap in time, the end-time value for the former Segment element is greater than the start-time value for the latter Segment. Still, within the bounds of each Segment, strict temporal linearity is maintained. The particular words in each Segment that are affected by the temporal overlap are delimited with hash-mark characters.

3.3. Transcription Methods

The tight scheduling for creation of transcripts, together with the problems imposed by having to start the transcription before the specifications were finalized, made it very difficult to establish an optimal set of tools and procedures for transcribers to use. The LDC created a transcription tool that combined the *xwaves* waveform editing utility from Entropics Research Labs, the *GNU Emacs* text editor, and a custom set of Tk/Tcl scripts and Emacs lisp functions. This combination of software elements provided transcribers with the ability to control the display, playback and marking of waveform regions; select attribute values (such as speaker name and speaking mode) for SGML tags; automatically insert correctly formatted tags, including correct time values from the waveform display; and type in transcription text – all using a flexible combination of Emacs keystrokes in the text buffer and mouse button selections on a graphical user interface. Knowing that the SGML format specifications were likely to change over the course of the transcription effort, we designed the custom modules to be easily reconfigurable.

Despite the power and flexibility of the transcription interface, the overall task faced by transcribers was excessively complex, especially as later developments in the transcription specification entailed changes in their practices. Although the interface helped to reduce the number and variety of typographic errors affecting the insertion of SGML tags, the task of correctly tagging so many different types and levels of information remained difficult, time-consuming, and error-prone. Numerous cycles of quality checking, correction, and reformatting were needed to produce documents that were reasonably accurate, usable, and compliant to the final specification [1].

The preparation of transcripts for the evaluation test sets was made considerably more tractable and reliable by breaking the task up into two stages. The first stage involved only the insertion of SGML tags to mark element boundaries and background changes; this was done on the entire 10-hour pool of material for a given test set. Based on the conditions indicated by these tags, NIST selected a subset of 2 hours worth of material from the pool, and just this subset was then submitted to careful correction of SGML tags followed by transcription.

4. Acknowledgements

I am indebted to other members of the LDC staff for their assistance in the preparation of this corpus: Rebecca Finch, Robert MacIntyre, Zhibiao Wu, and Mark Liberman. I would also like to express my thanks to George Doddington for his efforts in finalizing the transcription specifications, and to the entire speech group at NIST for their assistance throughout the project.

References

1. Doddington, G. “The 1996 Hub-4 Annotation Specification for Evaluation of Speech Recognition on Broadcast News,” *ftp://jaguar.ncsl.nist.gov/csr96/h4/h4annot.ps*